International Journal of Research in Advent Technology, Vol.6, No.5, May 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

An effective hybrid Fuzzy C-means and Fuzzy K-modes clustering algorithm for social media data

Akash Shrivastava[1], Dr. M. L. Garg[2]
Computer Science Engineering Department DIT University, Dehradun
akash.10may@gmail.com[1], dr.ml.garg@dituniversity.edu.in[2]

Abstract: Web space is flooded with unstructured data. The main reason behind this phenomenon is the usage of social media forum frequently. Every individual possibly creates 1 MB data on average. Data created over web in unstructured format may be useful and plays important role in decision making process undergoing in organization. Clustering analysis algorithms have been evolved from past years are now utilizing for categorizing social media data. Fuzzy C-means is one of the known clustering algorithms in the same direction. Fuzzy K-modes algorithm is also known to be underlying in same category. In this paper, these two algorithms is being integrated to mesh their merits and proposed hybrid fuzzy clustering approach implemented on social media data Twitter. The experiments executed are encouraging and preferable to utilized for categorical data.

Keywords: Big data; Twitter; Clustering; Artificial Bee Colony (ABC); Fuzzy clustering; Fuzzy C-means.

1. INTRODUCTION

In the current world, the enormous amount of social media data has blown the web space. The platform provided by the Social media becomes a tool to produce and analyze the reviews regarding any new technology, social issues and political perception including many other domains. In the context of social platform, Twitter and facebook both are known to be the master player which plays an important role in this direction from past recent years. Twitter becomes a routine and official forum where every popular individual across the globe or countrymen who has access to internet mostly marks their existence over twitter. In [1], it has been statically stated that 140 million plus active users used to publish 400 million 140 characters “Tweets” everyday. On the other hand facebook placed its mark as one of the prominent and includes highest userbase across the globe with 2 billion plus users. These both facts evolve the philosophy that what happen to the unstructured data it generated over the time. In fact in the recent time facebook faces the charge to leak the public data which is being misused for few political intentions. Now the scenario is that this social media takes analytics domain up to phenomenal heights where companies, organizations, institutions are willing to fetch the data produces over web space and used for their own commercial purposes. This willingness of the business and people actually initiated the association between clustering and big data analysis. Massive Unstructured data produced from twitter or facebook have been proven to be highly tangible to process and transformed into knowledge. Clustering analysis already known to be best data mining approach to cluster and group the data objects into similar and dissimilar groups [2]. In general, Clustering analysis considered as a well-developed and nurtured in four design phases: (i) Data representation phase which elaborates what kind of cluster structures can be identified in the data. (ii) Modeling phase lies when hidden structures involved in the data is produced. (iii) Optimization phase where quality measure has been performed which can be optimized up to certain level. (iv) Validation phase is the last phase which is responsible to validate and justify the results produced from the clustering algorithm. [3] [4] [5]

Earlier K-means is known to be the most popular clustering algorithm in the series of clustering algorithm. In case of real world data sets, K-means is proven to be ineffective where no definite boundaries lying between the clusters. Algorithms in clustering are classified into two classes: hard clustering algorithms and fuzzy clustering algorithms. In the first class, one object is only belonging to one cluster. In the latter case, every object belongs to every cluster by allowing having a membership function [6][7]. In the category of fuzzy clustering algorithms, Bezdek proposed [8] fuzzy C-means (FCM) is widely accepted and utilized for clustering purpose. Fuzzy C-means proven to be an effective and efficient algorithm for clustering but in case there is a requirement to select a random center points in cluster then it makes iterative process falling into the local optimal solution easily.

Fuzzy K-modes algorithm has developed as a result to improve the Fuzzy K-means algorithm as it only works on numeric values. [8] Fuzzy K-modes algorithm has been resolved the purpose of clustering categorical data sets like social media data. These both algorithms have been developed and tested on biological data. In this paper, their merits is being combined and applied over social media data which is highly categorical in nature.

The remaining of the paper is organized in the following way. In section 2, the related work has been investigated and section 3 introduces fuzzy C-means clustering. In section 4 fuzzy K-modes algorithm for clustering is discussed; Section 5 presents our hybrid clustering approach and Section 6 shows the experimental results. Final work is being concluded in Section 7.

2. RELATED WORK

Hesam and ajith [9] (paper 1) carried out the work over Fuzzy c-means algorithm and Particle swarm optimization (PSO) algorithm. They pointed out the merits of two algorithms and amalgamated as a hybrid efficient clustering algorithm. The experimental results which have been carried out over real data sets reveal encouraging results. The hybrid fuzzy clustering algorithm FCM-FPSO is proven to be a superior across fuzzy C-means and Particle swarm optimization algorithm. However, fuzzy C-means algorithm is highly sensitive to initialization and is observed to be easily falling into local optima. Unlike Fuzzy C-means, particle swarm optimization is being applied over various function optimization problems as a global tool. In due course of reducing the limitation of Fuzzy C-means algorithm it has been integrated with fuzzy Particle swarm optimization.

J. Wu and Z. Yang in [10] (paper 2) have introduced hybrid genetic fuzzy k-modes algorithm. They observed that clustering algorithm which falls into the local optima is not able to utilize for the categorical data sets. The work carried out in that context is recognized as the development of genetic fuzzy K-modes algorithm to overcome the limitation of local optima. Genetic algorithm and fuzzy K-modes algorithm has been integrated to find the global optimal solution to the optimization problem. The fuzzy clustering algorithm is being utilized as a solution for the optimization problem. They proposed the genetic fuzzy k-Modes algorithm for the purpose of clustering categorical data. Fuzzy K-modes have been treated as an optimization problem and Genetic algorithm is being integrated to provide the solution to achieve the global optimal solution.

. In [11], fuzzy clustering problem can be represented as an optimization problem mathematically as:

$$\min_{W,Z} F\,(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\propto}\, d(z_l, X_i)$$

Where,

n= number of objects in the data set

k= number of clusters

D= {$x_{1,}$ $x_{2,}$ x_2 ………., x_n} is a set of n objects, here each object is described by d attributes

Z = {$z_{1,}$ $z_{2,}$ $z_{3,}$ …………., z_k} is a set of k cluster centers

W is a weighing component where W= (w_{li}) is a k × n fuzzy membership matrix

d($z_{l,}$ x_i) = Distance between cluster center z_l and the x_i object

3. FUZZY C-MEANS ALGORITHM

Fuzzy c-means based on the concept where it partitions set of b objects o= {$o_{1,}$ $o_{2,}$ $o_{3,}$ …………., o_n } in Q^d dimensional space into c (1<c<b) here c refers as the

fuzzy clusters associated with $Y = \{y_1, y_2, ..., y_c\}$ cluster centers. Here, b is the number of data objects and c is the number of clusters which is described in fuzzy clustering of objects by a fuzzy matrix μ which contains b as rows and c as columns. μ_{ij} is considered to be an element lies in the matrix at position of ith row and jth column. This position of element actually signifies the degree of association or membership function of the ith object linked with the jth cluster. According to [8], fuzzy C-means is iterative and can be stated as below:

Fuzzy C-means:

Step 1: b must be selected as b>1, membership function values has been initialized as μ_{ij} where i= 1,2,3,…….. b; j=1,2,3……, c.

Step 2: Cluster centers $y_{j,}$ j=1,2,3……, c is computed in this step as

$$Z_j = \frac{\sum_{i=1}^{b} \mu_{ij}^{m} o_i}{\sum_{i=1}^{b} \mu_{ij}^{m}} \qquad (1)$$

Step 3: Euclidean distance d_{ij} is computed at this stage, 1,2,3,…….. b; j=1,2,3……, c by

$$d_{ij} = \| o_i - y_j \| \qquad (2)$$

Step 4: Membership function is updated as per following formula, here apply μ_{ij} where i= 1,2,3,…….. b; j=1,2,3……, c.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{b-1}}} \qquad (3)$$

Step 5: In case if till step 4 it is not being converged then go back to step 2.

The fact must be considered here regarding the FCM algorithm is that it is observed to be highly sensitive towards the initial values and it is also prone to be fall into local optima.

4. FUZZY K-MODES ALGORITHM

Fuzzy k-modes algorithm proposed by huang and Ng in [9]. Algorithm proposed can be implemented recursively as elaborated in [10].

Fuzzy K-modes: i is the maximum number of iterations

Step 1: Initial point of cluster center has been chosen as C_0 Є R^{mk} where k represents the clusters.

Step 2: The membership function P_0 is to be identified in a manner that cost function $F(P_{0,} C_0)$ is minimized;

Step 3: for t= 1 to i {

C_1 is being identified such that cost function $F(P_{0,} C_1)$ is minimized;

If $F(P_{0,}$ C1) found to be equal to $F(P_{0,} C_0)$, then stop;

Step 4: elseif {

Step 5: C_1 in a manner that the cost function $F(P_{1,} C_1)$ is minimized;

Step 5: If $F(P_{1,} C_1)$ found to be equal to $F(P_{0,} C_1)$, then stop;

Step 6: else

Step 7: $P_0 \Leftarrow P_1$;

Step 8: end if

Step 9: end if

Step 10: end for

5. HYBRID FUZZY C-MEANS AND K-MODES ALGORITHM FOR CLUSTERING SOCIAL MEDIA DATA

Step 1: The parameters of FCM and Fuzzy K-modes are being initialized i.e. b and initial cluster point C_0 has been chosen.

Step 2: Define $D = \{x_1, x_2, x_3, ………, x_n\}$ be a categorical data set with n objects.

Step 3: Define each of the object by d categorical attributes as $A = \{A_1, A_2, ……………………A_d\}$.

Step 4: FCM algorithm

4.1 Determine the cluster center for each attribute by calculating it using eq. (1)
4.2 For each categorical attribute, computer Euclidean distance $d_{ij,}$ i= 1,2,3,…….. b; j=1,2,3……, c using eq. (2)
4.3 For each attribute, Membership function μ_{ij} where i= 1,2,3,…….. b; j=1,2,3……, c is updated using eq. (3)
4.4 Calculate cost function F ($\mu_{ij,}$ C_0) for each attribute
4.5 If FCM terminating condition not met, go to step 4.

Step 5: Fuzzy K-modes algorithm

5.1 The membership function P_0 is to be identified in a manner that cost function F ($P_{0,}$ C_0) is minimized;

5.2 For each attribute,

for t= 1 to i {

C_1 is being identified such that cost function F ($P_{0,}$ C_1) is minimized;

If F($P_{0,}$ C1) found to be equal to F($P_{0,}$ C_0), then stop;

5.3 elseif {

5.4 C_1 in a manner that the cost function F ($P_{1,}$ C_1) is minimized;

5.5 For each attribute, If F ($P_{1,}$ C_1) found to be equal to F($P_{0,}$ C_1), then stop;

5.6 else

5.7 $P_0 \Leftarrow P_1$;

5.8 end if

5.9 end if

5.10 end for

5.11 if fuzzy k-modes terminating condition is not met, go to step 5.

Step 6: If FCM-Fuzzy K-modes condition is not met, go to step 4.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the performance of our proposed hybrid FCM and Fuzzy K-modes clustering algorithm is being experimented and evaluated. The proposed approach has been executed through various runs on TWITTER real time streaming categorical dataset. Yangs accuracy measure [11] has been adopted for this research work and the Rand Index [12] is implemented to assess the clustering results. In Yangs method, the definitions of accuracy (AC), precision (PR), and recall (RE) is given as follows:

$$AC= \frac{\sum_{i=1}^{k} a_i}{n} \tag{4}$$

$$PR= \frac{\sum_{i=1}^{k} \frac{a_i}{a_i + b_i}}{k} \tag{5}$$

$$RE= \frac{\sum_{i=1}^{k} \frac{a_i}{a_i + c_i}}{k} \tag{6}$$

Where a_i is the number of data objects that are correctly allocated to class C_i, b_i is the number of data objects that are incorrectly allocated to class C_i, c_i is the number of data objects that are incorrectly denied from class C_i, k is the total number of class contained in a dataset, and n is the total number of data objects in a dataset, and n is the total number of data objects in a dataset. In the above measures the AC has the same meaning as the clustering accuracy r defined in [13]. Given a dataset X={x_1,x_2…..,x_n} as well as two partitions of this dataset: Y={y_1,y_2,y_{t1}} and Y'={y_1',y_2'……….., y_{t2}'}, the Rand Index(RI) [14] is given by

$$RI = \frac{\sum_{i=1,j=2;i<j}^{n} \alpha_{ij}}{\left(\frac{n}{2}\right)} \tag{7}$$

Where

$$\alpha_{ij} = \begin{cases} 1, if\ there\ exist\ t\ and\ t'such\ that\ both\ x_i\ and\ x_j\ are\ in\ both\ y_t\ and\ y'_{t'}, \\ 1, if\ there\ exist\ t\ and\ t'such\ that\ x_i\ is\ in\ both\ y_t\ and\ y'_{t'} \\ while\ x_j\ is\ in\ neither\ y_t\ or\ y'_{t'}, \\ 0, otherwise \end{cases}$$

The RI is calculated by using the true clustering and the clustering obtained from a clustering algorithm. According to these measures, a better clustering result is indicated from the higher values of AC, PR, RE and RI. In the performance analysis, the proposed hybrid FCM and fuzzy K-modes algorithm, implemented on real time streaming twitter datasets. Then the clustering results of the proposed hybrid FCM and fuzzy K-modes algorithm is being compared with that of the other two algorithms i.e fuzzy C-means and fuzzy k-modes in terms of the best (Best), average (Avg.), and standard deviation of AC, PR, RE, and RI. All algorithms are implemented in python language and executed on intel core i7, 3.9 GHz, 32GB RAM computer system. In all experiments, the parameters of the proposed hybrid FCM and fuzzy K-modes algorithm are set as follows k=4, N =20, α=1.2. c1=c2=2.0, w=0.9. The terminating condition for FCM clustering algorithm is set to be 1.

Table 1. The AC of the three algorithms on the Twitter dataset

Algorithms	AC		
	Best	Avg	Std
FCM and Fuzzy K-modes	0.9236	0.9234	0.0209
Fuzzy C-means	0.9837	0.9887	0.0134
Fuzzy K-modes	0.9407	0.9233	0.0407

Table 2. The PR of the three algorithms on the Twitter dataset

Algorithms	PR		
	Best	Avg	Std
FCM and Fuzzy K-modes	0.8934	0.8778	0.0093
Fuzzy C-means	0.8912	0.8762	0.0096
Fuzzy K-modes	0.9184	0.8783	0.0137

Table 3. The RE of the three algorithms on the Twitter dataset

Algorithms	RE		
	Best	Avg	Std
FCM and Fuzzy K-modes	0.7881	0.8118	0.0078
Fuzzy C-means	0.7883	0.8127	0.0086
Fuzzy K-modes	0.8212	0.8132	0.0083

The Application Of the hybrid fuzzy clustering Algorithm on the Proposed twitter data sets shows an effective sign of improvement in the best ,average and lower standard values in AC, PR, RE and RI and therefore the algorithm considered to provide a better computability than the rest of the existing clustering algorithms used for comparison. The result set proven the implementation of algorithm on various data sets which are taken as classless and unorganized data set values.

6.1. Computation analysis

The computation analysis of proposed research work has been carried out to justify the objective of hybrid approach. Twitter data set is being experimented on three mentioned algorithms where our hybrid proposed algorithm which is the integration of fuzzy K-means and fuzzy c-means algorithm proven to be performed efficiently. The computation time have been obtained for the same where for all searches the computation time of proposed algorithm observed to be highest. It actually initiates the direction of research that supposed to apply over categorical data frequently encountered across social platforms. Here the computation has been done on profile IDs of different twitter handler and hence cluster analysis has been carried out by applying the proposed efficient hybrid algorithm.

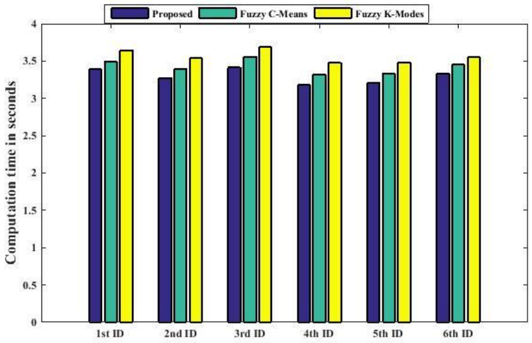

Figure 1. The computation time obtained for 6 Randomly chosen profile ID from experimented Twitter Dataset

6.2. Convergence analysis

In due support of our work, the convergence analysis has been done which justifies the approach of hybrid algorithm where the claim is being proven that if the proposed hybrid algorithm have been applied over categorical twitter dataset for every search of profile the performance goes up comparatively among other two algorithms. The cluster formation of tweets is very frequent in our approach. The four clusters have been observed and analyzed for the experiment carried out in the proposed research.

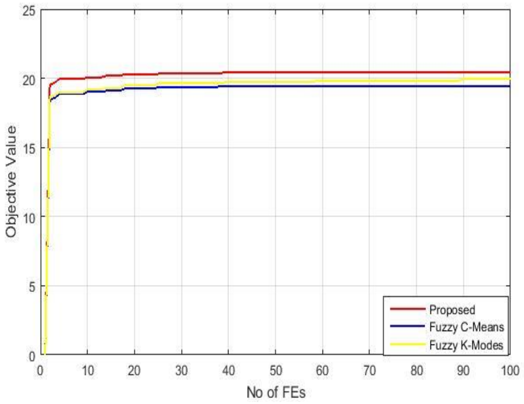

Figure 2. The Convergence plot of different optimization algorithm for four clusters

7. CONCLUSION

In this research we integrated the FCM and Fuzzy K-modes clustering approach which is the efficient form of existing clustering algorithm on basis of its merits. It has been observed that fuzzy K-Modes algorithm is based on the traditional clustering algorithm and FCM approach is highly sensitive towards the initial value. These both algorithms are prone to fall into local optima. In proposed hybrid clustering algorithm, the implementation is being carried out over categorical attributes by applying it on Twitter social media datasets. The potential of clustering algorithm has been observed in context of categorical social media data set. The experiment that has been tested actually proved the algorithm as highly effective. The experimental results demonstrated that the proposed algorithm was superior to other two well-known algorithms according to evaluation measures AC, PR, RE and RI respectively.

In near future, social media platforms like facebook, YouTube could be experimented by utilizing the proposed hybrid clustering algorithm to make clustering efficient for big data spreading across web. Furthermore, the hybrid approach would encourage the thought to develop the clustering algorithm which must be suitable for unstructured data with categorical attributes as well as mixed data.

REFERENCES

[1] Shamanth Kumar, Fred Morastatter, Huan Liu, Twitter Data analytics, Springer, Aug 19,2013.

[2] Cormack, R. (1971). A review of classification. Journal of the Royal Statistical Society. Series A (General), 134(3), 321–367.

[3] Gordon, A. (1987). A review of hierarchical classification. Journal of the Royal Statistical Society. Series A (General), 150(2), 119–137.

[4] Cowgill, M., Harvey, R., & Watson, L. (1999). A genetic algorithm approach to cluster analysis. Computers and Mathematics with Applications, 37(7), 99–108.

[5] Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. The Computer Journal, 26(4), 354–359.

[6] Everitt, B., Landau, S., & Leese, M. (2001). Cluster analysis (4th ed.). New York: Oxford University Press.

[7] Jain, A., & Dubes, R. (1988). Algorithms for clustering data. Englewood Cliffs, New Jersey: Prentice Hall.

[8] Bezdek, J. (1974). Fuzzy mathematics in pattern classification. Ph.D. thesis. Ithaca, NY: Cornell University.

[9] Huang, Z., & Ng, M. (1999). A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, 7(4), 446–452.

[10] Gan, G., Wu, J., & Yang, Z. (2009). A genetic fuzzy k-modes algorithm for clustering categorical data. Expert Systems with Applications (36), 1615–1620.

[11] Yang Y. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval. 1999; 1: 67–88.

[12] Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. 1971; 66: 846–850

[13] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values.

Data Mining and Knowledge Discovery. 1998; 2: 283–304.

[14] Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. 1971; 66: 846–850.